Egor Khairullin, Andrey Ustyuzhanin

Speeding up prediction performance of the BDT-based models.
Introduction

Online Trigger:

› Sample: one proton-proton collision
› Binary classification: is event interesting or not (B decay)
› Event consists of:
   1. secondary vertices (SV)
   2. tracks (track description)

http://arxiv.org/abs/1510.00572
Data features

- decay vertex of studied particle
- flight distance
- number of tracks
- beam axis (z-axis)
- transverse momentum (x, y)-plane
- mass, corrected mass, transverse momentum (x, y)-plane
- pseudorapidity
- SV impact parameter
- impact parameter of second track
- PV
- SV

Trigger system in ML terms
Machine learning problem

"Signal":
    simulated events for decays channels of interest

"Background":
    simulated generic proton-proton collisions

Goal:
    get the highest signal efficiency given background rejection rate
Figure of Merit

- Optimize Area under ROC curve in a region with small False Positive Rate ($<0.05$)
- Corresponds to mean efficiency of the model under given bandwidth limit
Restrictions

› Real-time event processing (time restriction, apply model to events one by one)

› Limited memory
Speeding up performance of the BDT

BDT & DTensor
BBDT (DecisionTensor)

- Convert decision trees to $n$-dimensional table (lookup table), $n$ is number of features
- Very fast
- Number of bins for each feature should be small (otherwise size grows too big)

$$D = (f_1, \ldots, f_n)$$
$$f_i = \{b_{i1}, \ldots, b_{ip_i}\}$$
$$S(D) \propto \prod |f_i| = \prod p_i$$
DecisionTensor Ensemble

› Several DecisionTensors with different bins
› Slower than one DecisionTensor
› Higher quality could be achieved
› How to choose ensemble parameters?

CatBoost & DTensor
DecisionTensor Similarity

\[ D^1 = (f_1^1, \ldots, f_n^1) \]
\[ D^2 = (f_1^2, \ldots, f_n^2) \]

\[ D^u = D^1 + D^2 \]
\[ D^u = (f_1^u, \ldots, f_f^u), f_i^u = f_i^1 \cup f_i^2 \]

\[ \text{Sim}(D^1, D^2) = S(D^1) + S(D^2) - S(D^1 + D^2) \]
Merging trees into DecisionTensor Ensemble

Given N, K:

1. Build BDT that consists of N trees
2. Initialize K empty DecisionTensors.
3. Take next tree T from BDT model, make new DecisionTensor from it and merge it with the most similar DecisionTensor.
4. Repeat 3. until all trees has been used.
CatBoost

› CatBoost is an open-source gradient boosting library (BDT) from Yandex
› Uses oblivious trees
› Discretize features
› Can improve existing model (baseline)
› Trained model usually consists of hundreds trees (and may be slow)
› https://github.com/catboost
CatBoost to DecisionTensor Ensemble (DTE)

1. Choose number of DTensors (K) by time restrictions;
2. Train CatBoost with few bins (memory restrictions) for each feature. Use it as Baseline for CatBoost models generated later;
3. Merge all trees into first DecisionTensor (D₁) of Ensemble;
4. Random search parameters for CatBoost;
   - Train CatBoost model
   - Convert model to DTE of size K-1
   - Remove DTEs exceeding memory restrictions
5. Choose DTE with the best quality on the test sample.
Experiment

Speeding up performance of the BDT
Setup

› Data

  Mean to recompute event prediction from SV predictions
  1M SVs, 160K events, 10 features
  Train - 50%, Test - 25%, Validation Set - 25%

› Up to 1 GB Memory

› Measure time of inference, applying model to each SV (one by one)
  of the whole Validation Set.
Lookups vs Time

CatBoost Best: 5000 trees
CatBoost Fast: 10-125 trees
DTE: 1-6 DTs
DTE size 150mb-1gb

Experiment
Quality vs Time

CatBoost Fast: 10-125 trees  
DTE: 1-6 DTs  
DTE size 150mb-1gb
DTE vs CatBoost

› Equal FOM:
  - DTE is up to 2.5 times faster and has up to 25 times fewer lookups

› "Best" CatBoost model (5000 trees):
  - DTE is 50 times faster and has 1000 less lookups
  - Difference in FOM only 0.006 (less than 1 per mil)

› Memory access is slow so DTensor lookups cost more than simple tree lookups (which can be cached)

› Faster memory (DDRAM5? FPGA?) could speed up DTensor lookups
Summary

DecisionTensor Ensemble

› helps running complex BDT (such as CatBoost) in real-time
› improves quality compared to single BBDT
› gives trade-off memory vs speed with marginal quality decrease
Thanks for attention!
Egor Khairullin
Lead software engineer

✉️ mikari@yandex-team.ru

Andrey Ustyuzhanin
Head of research group

✉️ anaderi@yandex-team.ru
## Comparison

<table>
<thead>
<tr>
<th>Model</th>
<th>ROC</th>
<th>Time, ms</th>
<th>Size, Mb</th>
</tr>
</thead>
<tbody>
<tr>
<td>catboost - 5000</td>
<td>0.6004</td>
<td>10000</td>
<td>&lt;5</td>
</tr>
<tr>
<td>catboost - 100</td>
<td>0.589</td>
<td>160</td>
<td>&lt;5</td>
</tr>
<tr>
<td>catboost - 75</td>
<td>0.5808</td>
<td>125</td>
<td>&lt;5</td>
</tr>
<tr>
<td>catboost - 10</td>
<td>0.5336</td>
<td>30</td>
<td>&lt;5</td>
</tr>
<tr>
<td>big dte - 6</td>
<td>0.5952</td>
<td>180</td>
<td>1000</td>
</tr>
<tr>
<td>big dte -3</td>
<td>0.5938</td>
<td>145</td>
<td>1000</td>
</tr>
<tr>
<td>big dte -2</td>
<td>0.5844</td>
<td>70</td>
<td>1000</td>
</tr>
<tr>
<td>small dte - 2</td>
<td>0.5794</td>
<td>55</td>
<td>150</td>
</tr>
<tr>
<td>dte - 1</td>
<td>0.5618</td>
<td>30</td>
<td>150</td>
</tr>
</tbody>
</table>
## Decays

<table>
<thead>
<tr>
<th>mode</th>
<th>2.5 kHz</th>
<th>4. kHz</th>
</tr>
</thead>
<tbody>
<tr>
<td>$B^0 \rightarrow K^*[K^+\pi^-]\mu^+\mu^-$</td>
<td>1.64</td>
<td>1.72</td>
</tr>
<tr>
<td>$B^+ \rightarrow \pi^+K^-K^+$</td>
<td>1.59</td>
<td>1.65</td>
</tr>
<tr>
<td>$B^0_s \rightarrow D_s^-[K^+K^-\pi^-]\mu^+\nu_\mu$</td>
<td>1.14</td>
<td>1.47</td>
</tr>
<tr>
<td>$B^0_s \rightarrow \psi(1S)[\mu^+\mu^-]K^+K^-\pi^+\pi^-$</td>
<td>1.62</td>
<td>1.71</td>
</tr>
<tr>
<td>$B^0 \rightarrow D_s^-[K^+K^-\pi^-]\pi^+$</td>
<td>1.46</td>
<td>1.52</td>
</tr>
<tr>
<td>$B^0 \rightarrow D^+[K^-\pi^+\pi^+]D^-[K^+\pi^-\pi^-]$</td>
<td>1.40</td>
<td>1.86</td>
</tr>
</tbody>
</table>
CatBoost to DecisionTensor Ensemble

› Choose count of DTensors by time restrictions.

› First DecisionTensor:
  
  Train CatBoost with few bins (memory restrictions) for each feature
  Merge all trees into first DecisionTensor ($D_1$) of Ensemble

› Next DecisionTensors:
  
  Train CatBoost with default bins but low tree count (due to memory and time restrictions)
  Merge trees to DecisionTensors ($D_2, D_3, ..., D_k$).
Models to compare

- CatBoost default model, 5000 trees
- CatBoost small models: 10-125 trees
- DT Ensemble: size from 1 to 6